Supplementary Note 1: Survey of previous and related concepts

1.1 Short introduction to multiple testing

In the multiple testing problem, we want to test m hypotheses $H_1, \ldots, H_m$ based on their corresponding p-values $P_1, \ldots, P_m$. From these, m_0 tests (with indices in $\mathscr{H}_0$, i.e. $|\mathscr{H}_0| = m_0$) are true nulls, while m_1 tests (with indices in $\mathscr{H}_1$) are alternatives. A multiple testing procedure will reject a certain fraction of the hypotheses, and the possible outcomes are summarized in Table S1.

	Not-rejected hypotheses	Rejected hypotheses	Total
True nulls	U	V	m_0
False nulls	T	S	m_1
	$m - R$	R	m

Supplementary Table S1: Outcomes of a multiple testing procedure applied to m hypothesis tests

The first approaches to multiple testing were concerned with controlling the Family-wise Error Rate (FWER $= \Pr[V \geq 1]$) at a pre-specified level α. In many applications, especially ones of exploratory nature, the FWER turns out to be too conservative. For instance, when screening many thousand substances, and detecting dozens of hits, one might be willing to accept a few false hits among them, as they may be eliminated later under a more detailed inspection. For this reason, an error measure which has gained significant popularity is the FDR (False Discovery Rate) [1], it is defined as the expected value of the FDP (False Discovery Proportion):

$$\mathrm{FDR} = \mathbb{E}[\mathrm{FDP}] = \mathbb{E}\left[\frac{V}{R \vee 1}\right] \tag{1}$$

The FDR has many advantageous features compared to existing error measures, for example it is adaptive, in the sense that the multiplicity penalty incurred adapts to the signal in the data and the number of tests [29]. Another reason for its widespread adoption is the existence of a simple and easy-to-use multiple-testing procedure which controls the FDR at a pre-specified level α. This procedure, named after Benjamini and Hochberg (BH) [1], works as follows: reject all hypotheses with p–values $P_i \leq \widehat{t}^*$, where $\widehat{t}^*$ is the solution of the following constrained maximization problem:

$$\text{maximize } R(t), \text{ s.t. } \frac{mt}{R(t)} \leq \alpha, \ t \in [0,1] \tag{2}$$

Here $R(t)$ denotes the number of p–values $\leq t$, that is, the number of rejections if t is used as the rejection threshold.
Beyond its simplicity (computation in just $\mathcal{O}(n \log n)$ operations; implementation in R's `stats:p.adjust` function takes just four simple lines of code) and the fact that it "always works" (proven FDR control under independence and diverse dependence structures [27]), the BH procedure also enjoys theoretical support through many different angles. In particular, while the FDR, as defined in (1), is a frequentist concept, the BH procedure as described by the constrained maximization (2) can be motivated in the Bayesian framework with the two-groups model [30, 31]. For this, assume that each H_i is a random variable with values in $\{0,1\}$, let $\pi_0 = \Pr[H_i = 0] \in [0,1]$ and let F_0, F_1 be the distribution functions corresponding to null and alternative hypotheses:

$$H_i \sim \text{Bernoulli}(1 - \pi_0) \tag{3}$$

$$P_i | H_i \sim (1 - H_i) F_0 + H_i F_1 \tag{4}$$

In other words, the p–value P_i has the marginal distribution

$$P_i \sim F = \pi_0 F_0 + (1 - \pi_0) F_1 \tag{5}$$

Then, the (Bayesian) False Discovery Rate (Fdr) of a multiple testing procedure rejecting all p–values $\leq t$ can be defined as:

$$\mathrm{Fdr}(t) = \frac{\pi_0 F_0(t)}{F(t)} \tag{6}$$

Note that for p–values, the standard assumption is that $F_0(t) = t$, i.e. the p–values are uniformly distributed under $\mathscr{H}_0$, thus:

$$\mathrm{Fdr}(t) = \frac{\pi_0 t}{F(t)} \tag{7}$$

This quantity now has a natural empirical estimator:

$$\widehat{\mathrm{Fdr}}(t) = \frac{\widehat{\pi_0}\, t}{\widehat{F}(t)} = \frac{\widehat{\pi_0}\, mt}{R(t)} \tag{8}$$

,
where $R(t) = \sum_{i=1}^{m} \mathbf{1}_{\{P_i \leq t\}}$, $\widehat{F}(t)$ is the ECDF and $\widehat{\pi_0}$ is an estimator of π_0.

If we conservatively estimate $\widehat{\pi_0} = 1$, then

$$\widehat{\mathrm{Fdr}}(t) = \frac{mt}{R(t)} \tag{9}$$

With the above considerations, the maximization in (2) just reads:

$$\text{maximize } R(t), \text{ s.t. } \widehat{\mathrm{Fdr}}(t) \leq \alpha,\ t \in [0, 1] \tag{10}$$

In other words, the BH procedure estimates the Bayesian Fdr for all rejection regions of the form $[0, t]$ simultaneously and then chooses a rejection region in the most greedy fashion. Thus, as observed by Efron [4], the BH procedure is so exciting because frequentist and Bayesian ideas coincide and because the conservativeness of the $\widehat{\mathrm{Fdr}}(t)$ estimator perfectly counteracts the greediness of the choice of rejection region.
Even though the BH procedure is more powerful than FWER-based methods, power can still be limiting in applications, and the multiple testing burden too high. Most of the theoretical advances have focused on estimating π_0, the proportion of null hypotheses, and incorporating that estimate into the multiple testing procedure. Unfortunately, this only provides a negligible increase in discoveries in real data sets if π_0 is close to 1, which is typically when power matters most. On the other hand, in practical data analysis, various heuristics have been developed in different applied fields (microarrays, eQTLs, mass spectrometry). Below we review these approaches with a particular emphasis on pointing out possible caveats when applying these.

1.2 Existing ways to increase power by use of covariates and a few notes of caution

1.2.1 Independent Filtering

Since the advent of microarrays, one of the most popular ways to increase power has been the filtering method [32, 33, 34]. Here, beyond the p–values $P_1, \ldots, P_m$, an additional covariate $X_1, \ldots, X_m$ is assumed to be available for each of the m tests. The procedure then works as follows: Let $x \in \mathbb{R}$ be a fixed threshold. In a first step, all hypotheses H_i with $X_i < x$ get filtered and then, in the second step, the classical BH procedure is applied to the remaining hypotheses.
This approach got heavily criticized for potentially leading to loss of FDR control [35]. On the other hand, it was generally believed to maintain FDR control

in cases where the covariate X is "non-specific". Later, Bourgon, Gentleman and Huber [9] provided some clarification. They first pointed out caveats and situations in which applying such a filter can actually cause loss of type-I error control, even when the covariate is "non-specific". Second, they provided a sufficient condition under which such a filtering method is valid: when the covariate is independent of the p–value under the null hypothesis ($P_i \perp X_i$, $i \in \mathscr{H}_0$), then the above two-step procedure controls the FDR at the pre-specified level ("Independent Filtering"). Multiple examples of such covariates were derived, and the potential for large increase in the number of discoveries was demonstrated.
However, [9] did not provide a rule or automatable method for the choice of the filter threshold x. The choice of this threshold has been criticized for being subjective [26]. What is more important though, is that researchers often do not realize the role of this parameter in type-I error control. In particular, control for a fixed choice of the threshold does not imply control over all thresholds simultaneously. Thus, allowing the researcher to set this parameter can lead to a problem similar to *p–value hacking* or *researcher degrees of freedom* [36]: even though the statistical produre, as reported in the published manuscript (with a fixed threshold reported) is valid, the actual validity is contingent on whether the researcher also tested different thresholds. To illustrate this further, we consider a Greedy Independent Filtering procedure, in which the researcher tests all possible thresholds and chooses the one which maximizes the number of discoveries.

Theorem 1 (Null case counterexample)**.** *Assume that we are performing m hypothesis tests based on p–values P_i and covariates X_i with P_i and X_i independent under the null hypothesis. All of these tests are null, i.e. $m = m_0$. Also assume that the corresponding p–values $P_1, \ldots, P_m \sim U[0,1]$ are i.i.d. Then the greedy Independent Filtering procedure does not control the* FDR *at level $\alpha \in (0,1)$, and the following lower bound holds:*

$$\text{FDR} \geq \sum_{i=1}^{m} \frac{\alpha}{i} \prod_{j=1}^{i-1} \left(1 - \frac{\alpha}{j}\right) > \alpha + \frac{\alpha}{2}(1-\alpha) > \alpha$$

This theoretical result is corroborated in the simulation in Figure 2e.

1.2.2 Stratified Benjamini–Hochberg (SBH)

Similarly to the Independent Filtering procedure, another method which has been often used in practical applications to increase power is the SBH (Stratified BH) procedure. The SBH procedure also uses an external covariate for each hypothesis test. In particular, the hypothesis tests are categorized into different strata based on this covariate. Then SBH proceeds as follows: The BH procedure is applied within each stratum at level α, and the rejections across the different strata are combined. The stratification arises naturally in the case of categorical covariates; in the continuous case the hypothesis tests are often binned according to increasing value of the covariate. For example, Degner *et al.* [37] split the eQTL hypotheses into 10 bins based on DNAseI sensitivity measurements and then applied Storey's q–value procedure within each bin [3].
Like Independent Filtering, the SBH approach can substantially increase discoveries [38]. In addition certain asymptotic justifications for the validity of the method have been provided [38, 28], while Efron [39] has shown the validity from an empirical Bayes / Bayes perspective. Despite these advantages of this approach, we would like to point out two caveats. First, in terms of controlling the global FDR (i. e., combined across the different strata), the SBH method can lead to loss of frequentist-FDR control in the case of $\pi_0 \approx 1$. This is also shown in Figure 2a.
The second caveat is related to power: the SBH procedure essentially controls the FDR within each stratum at level α. On the other hand, if global FDR control is of interest, then there is no a-priori reason to assign the same significance (type-I error budget) to each stratum. Indeed, to maximize power, different strata should be prioritized differentially [20].

1.2.3 Weighted Benjamini–Hochberg

A third general approach to increasing the power of multiple testing procedures is to weight each hypothesis according to the prospects of it actually showing a measurable true effect. Let $w_1, \ldots, w_m \geq 0$ be weights corresponding to the different hypotheses, such that $\sum_{i=1}^{m} w_i = m$. Now define $Q_i = \frac{P_i}{w_i}$ (with $Q_i = \infty$ for $w_i = 0, P_i \neq 0$ and $Q_i = 0$ for $w_i = 0, P_i = 0$). A weighted multiple testing procedure now operates on $Q_1, \ldots, Q_m$ rather than $P_1, \ldots, P_m$. Genovese et al. [6] showed that applying such a weighted BH procedure, i.e. applying the BH procedure to Q_i instead of P_i, provides finite sample FDR control in the case of independent p–values.
In fact, this weighted BH method can be seen as a generalization of both the Independent Filtering and SBH procedures. For the Independent Filtering case, assume that the filter threshold retains only $\widetilde{m}$ of the m hypotheses. Then we test the remaining p–values based on critical values $\alpha \frac{i}{\widetilde{m}}$, $i = 1, \ldots, \widetilde{m}$. This is equivalent to assigning weights $w_i = \frac{m}{\widetilde{m}}$ for the retained p–values and $w_i = 0$ for all other p–values and then applying the weighted BH procedure. For the SBH case, Yoo et al.[28] have shown that asymptotically, the SBH procedure is equivalent to assigning weights to the hypotheses in each stratum proportional to the number of rejections (of the BH procedure) in that stratum.
To illustrate how the weighted BH procedure (and other weighted multiple testing procedures [6]) can increase power, let t be a possibly data-driven threshold. Then, hypothesis H_i gets rejected if and only if $Q_i \leq t$ or equivalently if $P_i \leq w_i t$. In other words, hypotheses with $w_i > 1$ get rejected more easily and hence are prioritized. This prioritization can be motivated as follows: In the two-groups model (as described in Supplementary Note 1.1), it is assumed that all p–values follow the same distribution. In practice, it is more reasonable to assume that there is heterogeneity among the tests, so that each test has a different prior probability of being null ($\pi_{0,i}$) and a different alternative distribution ($F_{1,i}$). If $\pi_{0,i}, F_{1,i}, i \in \mathscr{H}$ were known, it would be possible to construct more powerful procedures than BH, which does not take this information into account.
Despite these advantages of the weighted BH procedure, in practice it has the same limitation as the Independent Filtering procedure (and even to a more extensive degree): It is not clear how to assign weights in a data-driven way, when oracle knowledge is not available. Given that (at least for the FWER-controlling procedures) robustness with regard to weights misspecification has been shown [40], many weighted procedures based on heuristic weight functions [41, 42, 43, 44] have been proposed. These do not generalize easily to other situations, possibly not even to other data sets of the same type.

1.2.4 Grouped Benjamini–Hochberg (GBH)

One general approach to make the derivation of data-driven weights for FDR control more tractable was suggested in [10], called the GBH (Grouped BH) procedure. Here, the authors assumed that the p–values were separated a-priori into G groups (strata). Rather than searching for an optimal weight for each hypothesis, all hypotheses in the same group are assigned the same weight and the specification of only G weights is required. A quasi-optimal heuristic for this assignment was proposed:

$$w_g \propto \frac{\widehat{\pi}_{1,g}}{\widehat{\pi}_{0,g}} \tag{11}$$

Here $\widehat{\pi}_{0,g}$ denotes the estimated proportion of null hypotheses in group g and $\widehat{\pi}_{1,g} = 1 - \widehat{\pi}_{0,g}$. As a π_0 estimator, the TST estimator [2] and the LSL estimator [45] were suggested.
Here we also need to point out two caveats. First of all, with the TST estimator, the FDR is again not controlled at a pre-specified level α in the $\pi_0 = 1$ case (Figure 2e). Also, the heuristic, quasi-optimal weight assignment can actually

lead to loss of power, even compared to the BH procedure, when the distribution of the alternatives differs across strata (Supplementary Fig. 2d,f).
Zhao et al. [46] aimed to improve the GBH method by starting with an application of BH and GBH and then assigning data-driven weights at fixed thresholds. This is achieved using an approach similar to our IHW-naive approach: The number of rejections is maximized subject to a fixed threshold (maximum of GBH and BH thresholds), rather than at a constrained plugin FDR value. No implementation is available though for the underlying optimization, and the choice of the GBH and BH thresholds is arbitrary.

1.3 Local fdr based approaches

In the Introduction it was mentioned that most theoretical work on increasing power has focused on π_0 estimation, while practical applications have used methods based on external covariates. Nevertheless, the idea of including covariates to increase power in multiple testing has been a lot more prevalent in the context of the local false discovery rate (fdr). This quantity is a local analog of the Bayesian Fdr (6) and is defined as follows (if the Lebesgue densities f_0, f_1 and f of F_0, F_1 and F exist):

$$\mathrm{fdr}(t) = \frac{\pi_0 f_0(t)}{f(t)} \tag{12}$$

In the stratified case, based on local fdrs, Ochoa et al.[21] considered the following constrained optimization problem under oracle knowledge:

$$\text{maximize } \mathbb{E}\left[R(t_1,\ldots,t_G)\right], \text{ s.t. } \mathrm{Fdr}(t_1,\ldots,t_G) = \frac{\mathbb{E}[V(t_1,\ldots,t_G)]}{\mathbb{E}[R(t_1,\ldots,t_G)]} \leq \alpha \tag{13}$$

$R(t_1,\ldots,t_G)$ denotes the number of rejections of a procedure which rejects all p-values in group g which are less than or equal to t_g. Similarly $V(t_1,\ldots,t_g)$ denotes the false rejections of such a procedure.
Now let Cfdr^g be the local fdr function conditional on belonging to stratum g. Then a necessary condition for maximization of (13) (in the interior of $[0,1]^G$) is that $\mathrm{Cfdr}^g(t_g) = \mathrm{Cfdr}^j(t_j)\ \forall\ j,g\ \in\ \{1,\ldots,G\}$ [21]. Based on this result, every optimal procedure is of the form: reject hypothesis H_i which belongs to stratum g if $\mathrm{Cfdr}^g(P_i) \leq q$ for a common, overall choice of $q \in [0,1]$. To achieve global FDR control at level α, q has to be chosen so as to fulfill the constraint $\mathrm{FDR} \leq \alpha$. Cai and Sun [20] considered exactly such multiple testing procedures and also showed that they are optimal in terms of minimizing the False Nondiscovery Rate (Fndr) subject to Fdr control (where Fndr and Fdr take the Bayesian definition). In addition, they provided a data-driven way of applying this oracle multiple testing procedure, which works as follows:
First, estimate the conditional local fdrs $\widehat{\mathrm{Cfdr}^g}(P_i)$ for all hypotheses H_i in stratum g, pool these estimates over all strata together and form their order statistics $\widehat{\mathrm{Cfdr}}_{(i)}$. Also let

$$j = \max\{i \mid \frac{1}{i}\sum_{i=1}^{m} \widehat{\mathrm{Cfdr}}_{(i)} \leq \alpha\}. \tag{14}$$

Then all hypotheses corresponding to $\widehat{\mathrm{Cfdr}}_{(1)},\ldots,\widehat{\mathrm{Cfdr}}_{(j)}$ get rejected.
A lot of further local fdr research has attempted to incorporate continuous covariates. One approach, which is interesting due to similarity to ours is the Covmod method [24], which also uses a covariate independent of the p–values under $\mathscr{H}_0$ to stratify the hypotheses. The authors make a parametric assumptions regarding the form of the distribution of the p–values within each bin: it is assumed to be a Beta-Uniform mixture. Fitting is done using an approximate Bayesian method, and information about parameters is shared across bins by choice of appropriate

priors, for example, for the $\pi_{0,g}$ (proportion of null hypotheses in the g-th bin):

$$f(\pi_{0,1}, \ldots, \pi_{0,G}) \propto \exp\left(\frac{-\lambda}{2} \sum_{g=2}^{G} \left(\text{logit}(\pi_{0,g}) - \text{logit}(\pi_{0,(g-1)})\right)^2\right) \tag{15}$$

with $\text{logit}(p) = \log\left(\frac{p}{1-p}\right)$ and a regularization parameter $\lambda > 0$.

A heuristic choice is suggested for the regularization parameter, but the user also has to manually choose a tuning parameter. Unfortunately, in our hands the available implementation of Covmod was not numerically stable and aborted with segmentation faults in many cases.
On the other hand, some approaches have been developed which do not rely on binning. A seminal contribution was made by Ploner et al. [22], whose key idea was to generalize the local fdr notion to higher dimensions (e. g., 2 dimensions) by:

$$\text{fdr2d}(z_1, z_2) = \pi_0 \frac{f_0(z_1, z_2)}{f(z_1, z_2)} \tag{16}$$

Nevertheless, the authors acknowledged the fact that routine application of fdr2d is limited by the difficulty of non-parametric ratio smoothing in higher dimensions. No automatic choice of smoothing parameter was provided, and the user was encouraged to inspect certain diagnostic plots.
Further work in this direction makes estimation possible by imposing additional parametric assumptions (just as the Covmod method). For example, the FDR-regression method [23] assumes that all hypotheses have the same alternative distribution and a Gaussian error model, while a logistic link with the covariate is assumed. The method of Zablocki et al. [47] assumes that the absolute z-scores follow a folded Normal-Gamma mixture and even more assumptions are imposed onto the covariate. Beyond the limitation of these parametric assumptions, the methods also suffer from loss of FDR control. For example, Zablocki et al. [47] point out the overoptimism when $\pi_0 \approx 1$.

1.4 Single-index modulated multiple testing

Du et al.[26] give another partial solution to the problem, by assuming that the covariate X is also a p–value. Thus, instead of a p–value and filter-statistic tuple (P, X), they consider a bivariate p–value (P^1, P^2). Given an angle $\theta \in [0, \frac{\pi}{2}]$, this bivariate p–value is projected onto a single index $P(\theta)$ defined by:

$$P(\theta) = \Phi(\cos(\theta)\Phi^{-1}(P^1) + \sin(\theta)\Phi^{-1}(P^2)), \tag{17}$$

where Φ denotes the standard normal cumulative distribution function.
The authors drop the assumption of conditional independence of the two p–values under $\mathscr{H}_0$ and instead assume that the bivariate p–value distribution is symmetric. Next, for fixed θ, π_0 and the null distribution of $P(\theta)$ are estimated non-parametrically (or parametrically under stronger assumptions) and a BH-type procedure is applied. The projection direction θ is calculated by applying the above procedure for each θ and choosing the one that maximizes the number of rejections. Under certain conditions (most importantly $\pi_0 < 1$), asymptotic FDR control is proven.
This method is similar to IHW-naive since the data-driven choice of the parameter θ is achieved by maximizing the number of rejections.

Supplementary Note 2: Full description of the IHW algorithm

2.1 IHW-naive

2.1.1 Algorithm

To derive the method, we assume that we have access to p–values $P = (P_1, \ldots, P_m)$ and a covariate $X = (X_1, \ldots, X_m)$, which is independent of P under $\mathscr{H}_0$. Our starting point for an algorithm (what in the paper is called the "naive algorithm") for data-driven choice of weights consists of the following steps:

- Group the hypothesis tests into G bins based on the values of the covariate X. Denote by m_g the number of hypotheses in the g-th bin, so that $\sum_{g=1}^{G} m_g = m$.
- For each possible weight vector $\mathbf{w} = (w_1, \ldots, w_G)$, i. e., for each $\mathbf{w}$ that satisfies $w_i \geq 0$ and $\sum_{g=1}^{m} w_g m_g = m$, apply the group-weighted BH procedure with this vector and calculate the total number of rejections. Then choose the[1] vector $\mathbf{w}^*$ that maximizes the number of rejections.
- Report the result of the grouped-weighted BH procedure with the optimal weight vector $\mathbf{w}^*$.

2.1.2 The idea behind the algorithm

This is essentially a greedy generalization of the BH procedure, as formulated above in terms of a constrained maximization (Equation (10)). In particular, here the plugin-FDR estimator introduced in Equation (9) corresponds to

$$\widehat{\mathrm{Fdr}}(t, \mathbf{w}) = \frac{mt}{R(t, \mathbf{w})} = \frac{\sum_{g=1}^{G} m_g w_g t}{R(t, \mathbf{w})} \tag{18}$$

We have used the notation $R(t, \mathbf{w})$ for the number of rejections of the projedure that rejects all hypotheses below $t_g := w_g t$ in bin g. Then the constrained maximization takes the following form:

$$\text{maximize } R(t, \mathbf{w}), \text{ s.t. } \widehat{\mathrm{Fdr}}(t, \mathbf{w}) \leq \alpha \tag{19}$$

We can also reparametrize the problem by $t_g = w_g t$, so that the hypotheses in each stratum get rejected based on a different thresholds. Then this procedure can be described as follows:

$$\widehat{\mathrm{Fdr}}(t_1, \ldots, t_G) = \frac{\sum_{g=1}^{G} m_g t_g}{R(t_1, \ldots, t_G)} \tag{20}$$

$$\text{maximize } R(t_1, \ldots, t_G), \text{ s.t. } \widehat{\mathrm{Fdr}}(t_1, \ldots, t_G) \leq \alpha, \ t_g \in [0, 1] \tag{21}$$

Note that this is just the maximization problem (13), which was solved in [21] in terms of the local fdr, expressed using the empirical quantities replacing the expected values (and upper-bounding the proportion of null hypotheses by 1).
Finally note that above we started from a formulation in terms of a weight vector $\mathbf{w}$ and a threshold t and derived the method in terms of individual thresholds t_g. This procedure can also be reversed, i.e. we can derive $\mathbf{w}$ and t from t_g. For this, first notice that from

$$t_g = w_g t \quad \text{and} \quad \sum_{g=1}^{G} m_g w_g = m \tag{22}$$

it follows that

$$\sum_{g=1}^{G} m_g t_g = \sum_{g=1}^{G} m_g w_g t = mt \tag{23}$$

[1] If the maximum is not unique, then just choose one of the optimal $\mathbf{w}$ randomly.

Now if $t_g = 0 \; \forall \; g \in \{1, \ldots, G\}$, then $t = 0$ and we can pick an arbitrary weight vector, e. g., $w_g = 1$. Otherwise $t > 0$ and the thresholds can be converted back to weights by

$$w_g = \frac{t_g}{t} = \frac{m t_g}{\sum_{g=1}^{G} m_g t_g} \tag{24}$$

2.2 IHW

We modify naive IHW to derive a procedure that is more scalable and has better finite sample properties, in 3 steps E1-E3.

2.2.1 Modification E1

Let $\widehat{F}_g$ be the ECDF of the p-values in stratum g. Then it holds that $R_g(t) = m_g \widehat{F}_g(t) \; \forall \; t \in [0, 1]$. Given that $R(t_1, \ldots, t_G) = \sum_{g=1}^{G} R_g(t_g) = \sum_{g=1}^{G} m_g \widehat{F}_g(t_g)$, we can rewrite equations (20) and (21) as follows:

$$\widehat{\mathrm{Fdr}}(t_1, \ldots, t_G) = \frac{\sum_{g=1}^{G} m_g t_g}{\sum_{g=1}^{G} m_g \widehat{F}_g(t_g)} \tag{25}$$

$$\text{maximize} \sum_{g=1}^{G} m_g \widehat{F}_g(t_g), \text{ s.t. } \widehat{\mathrm{Fdr}}(t_1, \ldots, t_G) \leq \alpha, \; t_g \in [0, 1] \tag{26}$$

From a more theoretical point of view, note that $\widehat{F}_g$ is just an estimator of the distribution function F_g in stratum g. In other words, at least heuristically we expect that our multiple testing procedure approximates the oracle procedure in which F_g is known. The asymptotic meaning of this will be made precise in our proof of asymptotic consistency of IHW in Supplementary Note 7. It is also instructive to compare to the oracle formulation (13) considered in Ochoa et al.[21].

For our first modification (E1), rather than estimating F_g using the ECDF, we use the Grenander estimator $\widetilde{F_g}$, i. e., the least concave majorant of the ECDF $\widehat{F_g}$. This can be calculated efficiently by applying the pooled adjacent violator algorithm (PAVA) in $\mathcal{O}(m_g \log(m_g))$ time for each stratum. Thus we impose the assumption that the distribution functions F_g are concave (i. e., have a decreasing density); this is a common and reasonable assumption in multiple testing [5].

This yields the following optimization problem:

$$\widetilde{\mathrm{Fdr}}(t_1, \ldots, t_G) = \frac{\sum_{g=1}^{G} m_g t_g}{\sum_{g=1}^{G} m_g \widetilde{F}_g(t_g)} \tag{27}$$

$$\text{maximize} \sum_{g=1}^{G} m_g \widetilde{F}_g(t_g), \text{ s.t. } \widetilde{\mathrm{Fdr}}(t_1, \ldots, t_G) \leq \alpha, \; t_g \in [0, 1] \tag{28}$$

Elementary reformulations, allow us to express this problem as follows:

$$\begin{aligned} \text{minimize} \;\; & H(t) = -\sum_{g=1}^{G} m_g \widetilde{F_g}(t_g) \quad \text{s.t.} \\ & H_1(t) = \sum_{g=1}^{G} m_g (t_g - \alpha \widetilde{F_g}(t_g)) \leq 0 \\ & t_g \in [0, 1] \end{aligned} \tag{29}$$

Note that H and H_1 are convex functions since $\widetilde{F_g}$ are concave, thus the optimization problem is convex.

Once we have these thresholds t_g, we can recover the weights w_g by (24) and apply the weighted Benjamini-Hochberg procedure.

Remark 1*:* It was pointed out to us, that our procedure could be too conservative, because the Grenander estimator of the distribution function often overestimates its values close to 0. We want to clarify here that the Grenander estimator only flows into the estimation of the weights; for the final p-value adjustment, the ECDF is used, since we just apply the standard weighted Benjamini-Hochberg procedure.

To make this more precise, recall the IH related optimization scheme (where this time we use a parametrization in terms of weights):

- For IHW naive, as in equation (26), we maximize over $t \in [0,1]$, $\mathbf{w}$ weight:

$$\sum_{g=1}^{G} m_g \widehat{F}_g(w_g t) \text{ s.t. } \widehat{\text{Fdr}}(w_1 t, \ldots, w_g t) \leq \alpha$$

- For IHW with Grenander, as in equation (28), we maximize over $t \in [0,1]$, $\mathbf{w}$ weight:

$$\sum_{g=1}^{G} m_g \widetilde{F}_g(w_g t) \text{ s.t. } \widehat{Fdr}(w_1 t, \ldots, w_g t) \leq \alpha$$

Now IHW with (E1) (and without modifications E2, E3), is actually neither of these! It is a 2-step approximation to solving (26) with the help of (28). It works as follows:

- Step 1: Solve (28) - which can be done efficiently since it is a convex problem - and get $\widetilde{\mathbf{w}}$ and $\tilde{t}$, which are solutions to the optimization problem. $\tilde{t}$ will not be needed in the following.
- Step 2: Solve (26) with $\mathbf{w}$ fixed to $\widetilde{\mathbf{w}}$. This yields $\hat{t}$ and can be solved e.g. by applying BH to the weighted p-values with weights $\widetilde{\mathbf{w}}$. Thus $(\hat{t}, \widetilde{\mathbf{w}})$ is used as the approximate solution to (26), which is a feasible point and defines the rejection thresholds.

In summary: (28) is used only to learn $\widetilde{\mathbf{w}}$, while weighted BH is used to learn $\hat{t}$. Therefore, biases due to the Grenander estimator might influence the weights, but not the final FDR estimator, which is still based on the ECDF! This allows us to combine computational efficiency, while also retaining the statistical properties of the ECDF, as it pertains to FDR controlling procedures.

2.2.2 Modification E2

We randomly split the hypotheses into n_{folds} folds. Splitting is done randomly, independently of the p-values and covariates of the individual tests. For each fold we proceed as follows:
We apply the optimization problem (29) to the hypothesis tests of the remaining $n_{\text{folds}} - 1$ folds. This yields a weight vector $\widetilde{\mathbf{w}} = (\widetilde{w}_1, \ldots, \widetilde{w}_G)$. Hypotheses which lie in stratum g of the held-out fold are then assigned weight $\widetilde{w}_g$.
Thus the hypotheses in each fold get assigned their weights. In total, we obtain a $n_{\text{folds}} \times G$ table of weights: one weight for each combination of fold and bin.
The key idea for modification E2 is that under our setting we can assume exchangeability of the p-value, covariate pairs (P_i, X_i). This is a much milder assumption than exchangeability of the p-values P_i. If this type of exchangeability holds, then we should be able to recover a good approximation to the optimal weight function by optimizing over a disjoint set of hypotheses; thus preventing "overfitting".
In particular, note that with E2 the weight assigned to a hypothesis does not directly depend on its p-value, but only on its covariate and on the p-values in the "training" folds. If the hypothesis tests are independent of each other, then the p-value P_i is independent of its assigned weight w_i under the null hypothesis, because the covariate is independent of the p-value under the null hypothesis. This is made more precise in Supplementary Note 6.2, where we prove that a variant of IHW, IHW-Bonferroni, controls the FWER.

2.2.3 Modification E3

Modification E3 adds further constraints to (29) such that the learning of the weight function can be improved and thus that the weights learned with $n_{\text{folds}} - 1$ folds generalize to the held-out fold.
To regularize our problem for ordered covariates, we add the additional constraint ("total variation" penalty) for a regularization parameter $\lambda \geq 0$:

$$\sum_{g=2}^{G} |w_g - w_{g-1}| \leq \lambda \tag{30}$$

$$\Rightarrow m \sum_{g=2}^{G} |t_g - t_{g-1}| \leq \lambda \sum_{g=1}^{G} m_g t_g \tag{31}$$

This constraint imposes that successive strata should not be too different. Adding (31) to optimization problem (29) maintains convexity of the program. Note that $\lambda = 0$ yields uniform weights, while $\lambda \to \infty$ corresponds to the unconstrained version. We will denote by IHW(λ) the IHW procedure with modifications E1, E2 and total variation constraint (30), so that IHW(∞) is IHW with E1 and E2.
In many situations we would like to determine a suitable value of λ from the data (a model selection problem). To ensure that P_i will still be independent of its assigned weight w_i under the null hypothesis (E2), this means that we have to learn the regularization parameter for each training set of $n_{\text{folds}} - 1$ folds individually. We then apply a nested cross validation step, which proceeds as follows for each training split:
We specify a finite grid Λ of regularization parameters. For each value $\lambda \in \Lambda$ we apply IHW(λ) to the training set hypotheses. In other words, the training set hypotheses get further split randomly into $n_{\text{folds,CV}}$ folds and (E2) is applied. We then choose the λ which led to the maximum number of rejections in the training set. In this situation, we can repeat the above with different random splits ($n_{\text{splits,CV}}$ splits) and choose the λ which led to the maximum number of rejections on average in the training set.
For an unordered covariate we can proceed in exactly the same way but instead use the constraint $\sum_{g=1}^{G} |w_g - 1| \leq \lambda$. This penalizes deviations from uniform weights.

2.3 Optimization

Here we describe how the optimization tasks (21) and (29), possibly with constraint (31), can be solved using a Mixed-Integer Linear Programming (MILP) solver (for IHW-naive) or a Linear Programming solver respectively (for IHW).

2.3.1 MILP optimization without regularization

Starting from the maximization problem (21), we can equivalently describe it by:

$$\begin{aligned} \text{maximize} \sum_{g=1}^{G} R_g(t_g) \;\; s.t. \\ t_g \in [0,1] \\ \alpha \sum_{g=1}^{G} R_g(t_g) \geq \sum_{g=1}^{G} m_g t_g \end{aligned}$$

Similar to the proof in the original BH paper [1], we observe that $R_g(\cdot)$ only changes its value at $P_i^g, i \in \{1, \ldots, m_g\}$ (the p-values in bin g). Thus it is enough to restrict our attention to these values of t (and $t = 0$).
For $g = 1, \ldots, G$, we have $(m_g + 1)$ ordered values (group-wise order statistics of the p–values):

$$0 =: P_{(0)}^g \leq P_{(1)}^g \leq \ldots \leq P_{(m_g)}^g \tag{32}$$

Since $R_g(P^g_{(i)}) = i$ almost surely, we get the following equivalent characterization (i.e. we write t as a function of R_g instead of the other way around)

$$\begin{aligned} &\text{maximize } \sum_{g=1}^{G} R_g \;\; s.t. \\ &R_g \in \{0, \ldots, m_g\} \\ &\alpha \sum_{g=1}^{G} R_g \geq \sum_{g=1}^{G} m_g P^g_{(R_g)} \end{aligned}$$

To express this problem as a MILP, we introduce the binary variables z^g_j, $j \in \{1, \ldots, m_g\}, g \in \{1, \ldots, G\}$ which satisfy the linear constraints:

$$z^g_1 \geq z^g_2 \geq \ldots \geq z^g_{m_g}$$

In the final formulation it will hold that $R_g = \sum_{j=1}^{m_g} z^g_j$, since the decision (binary) variable z^g_j denotes whether the j-th lowest p–value in group g will be rejected. Also define $y^g_k = P^g_{(k)} - P^g_{(k-1)}, k = 1, \ldots, m_g$. Then we can solve the equivalent MILP:

$$\begin{aligned} &\text{maximize } \sum_{g=1}^{G} \sum_{j=1}^{m_g} z^g_j \;\; s.t. \\ &z^g_j \geq z^g_2 \geq \ldots \geq z^g_{m_g} \in \{0,1\},\ g \in \{1, \ldots, G\} \\ &\alpha \sum_{i=1}^{G} \sum_{j=1}^{m_g} z^g_j \geq \sum_{g=1}^{G} \sum_{j=1}^{m_g} m_g y^g_k z^g_j \end{aligned}$$

2.3.2 MILP optimization with regularization

To enforce the constraint (31) we also have to introduce new variables t_g with $t_g \geq \sum_{j=1}^{m_g} y^g_k z^g_j$ and reformulate the FDR bound in terms of these new thresholds t_g. The final problem then has the following form:

$$\begin{aligned} &\text{maximize } \sum_{g=1}^{G} \sum_{j=1}^{m_g} z^g_j \;\; s.t. \\ &z^g_j \geq z^g_2 \geq \ldots \geq z^g_{m_g} \in \{0,1\},\ g \in \{1, \ldots, G\} \\ &t_g \geq \sum_{j=1}^{m_g} y^g_k z^g_j,\ g \in \{1, \ldots, G\} \\ &m \sum_{g=2}^{G} |t_g - t_{g-1}| \leq \lambda \sum_{g=1}^{G} m_g t_g \\ &\alpha \sum_{i=1}^{G} \sum_{j=1}^{m_g} z^g_j \geq \sum_{g=1}^{G} m_g t_g \end{aligned}$$

2.3.3 LP optimization without regularization

To solve problem (29), we observe that the Grenander estimator $\widetilde{F_g}$ is a piecewise-linear, concave function. There exists a finite index set I_g and real numbers a^g_i, b^g_i for $i \in I_G$ such that:

$$\widetilde{F_g}(t) = \min_{i \in I_g} \{a^g_i + b^g_i\, t\} \tag{33}$$

For each $t_g \in [0,1]$ we introduce a new variable $f_g \in [0,1]$ $(g = 1, \ldots, G)$ and add the constraints $f_g \leq a^g_i + b^g_i\, t_g$ for $i \in I_g$. Then we just need to maximize the linear function $\sum_{g=1}^{G} m_g f_g$ under these linear constraints and the plugin-FDR control constraint $\sum_{g=1}^{G} m_g (t_g - \alpha f_g) \leq 0$. This is a linear program.

2.3.4 LP optimization with regularization

The constraints (31) can be added directly to the linear program of the formulation with the Grenander estimator using standard methods for modelling absolute values in linear programming.

2.4 Choosing the number of bins

The number of bins depends on two factors: First, within each bin there should be enough p-values so that the distribution function can be estimated well by its Grenander estimator. In practice we recommend having at least 1000 p-values within each bin. Second, the difficulty of optimization problem (29) depends on the number of bins.

Supplementary Note 3: Real-data examples

3.1 DESeq2 (Bottomly) example

For the RNA-Seq example we used the dataset of Bottomly *et al.* [13], which we downloaded from the Recount project [14]. p-values were calculated using DESeq2 [12] with default settings, for the design $\sim$ cell + dex. We used the mean of normalized counts for each gene, across samples, as the informative covariate. Hypotheses were stratified into 13 equally sized bins (i. e., with the same number of hypotheses) based on the covariate. IHW was used with settings $n_{\text{folds}} = n_{\text{folds,CV}} = n_{\text{splits,CV}} = 5$. This was repeated for nominal levels $\alpha \in [0.05, 0.1]$ using an equidistant grid with 5 values.
We compared the result to that of using the Benjamini–Hochberg method.

3.2 Proteomics (Gygi) example

We used the dataset in [15] (from their Supplementary Table 1), in particular, their Welch t-test p-values, and the number of peptides quantified as the covariate. Hypotheses were stratified into 4 equally sized bins (same number of hypotheses) based on the covariate. IHW was used with parameters $n_{\text{folds}} = n_{\text{folds,CV}} = n_{\text{splits,CV}} = 5$, and the regularization parameter was selected from a grid of 20 equidistant values in $\lambda \in [0, 3]$. This was repeated for nominal levels $\alpha \in [0.05, 0.1]$ (equidistant grid with 5 values).
We compared the result to that of using the Benjamini–Hochberg method.

3.3 hQTL example

For the hQTL example, we used the dataset described in [16] and looked for associations between SNPs and the histone modification mark (H3K27ac) on human Chromosome 21. p-values for association were calculated as described in the original paper [16] using Matrix eQTL [48]. As a covariate we used the linear genomic distance between the SNP and the ChIP-seq signal.
We stratified hypotheses based on the distance in 10 kb bins up to 300 kb, 100 kb bins up to 1 Mb, 10 Mb bins for the rest of the hypotheses. We used IHW with $n_{\text{folds}} = 5$ and without the E3 step, i. e., we set $\lambda = \infty$. This was repeated for nominal levels $\alpha \in [0.05, 0.1]$ (equidistant grid with 5 values).
We compared the result to Benjamini–Hochberg and Independent Filtering with thresholds set to 10 kb, 200 kb and 1 Mb.

Supplementary Note 4: Simulation studies

4.1 Implementation of benchmarked methods

4.1.1 IHW-naive

For IHW-naive we stratified all hypotheses into 20 bins of equal size. To solve the underlying MILP problem we used the Gurobi solver version 6.5 [49].

4.1.2 IHW

For IHW we also stratified all hypotheses into 20 bins of equal size. For the underlying LP problem we used the open-source SYMPHONY solver of the COIN-OR project [50]. A total variation penalty of the weights was used with the regularization grid $\lambda \in \{0, 1, 2.5, 5, 10, 20, \infty\}$ and $n_{\text{folds}} = 5$, $n_{\text{folds,CV}} = 5$, $n_{\text{splits,CV}} = 1$.

4.1.3 BH

For BH we used the standard `p.adjust` function with method `BH` in R's `stats` package.

4.1.4 Greedy Independent Filtering

We applied the Greedy Independent Filtering procedure by applying BH for each of the m relevant filter statistic thresholds and choosing the threshold that maximized rejections. Then we applied the independent filtering procedure with that threshold [9].

4.1.5 LSL-GBH, TST-GBH

The Group Benjamini–Hochberg (GBH) [10] procedure requires stratification of the hypotheses into bins (i. e., categorical covariates). For the simulations we stratified into 20 bins of equal size based on increasing value of the covariate. The LSL-GBH and TST-GBH procedures were implemented as described in the publication [10].

4.1.6 SBH

For the stratified BH procedure, hypotheses were stratified as for the GBH procedure. Within each stratum the BH procedure was applied at level α, and the rejections across all strata were pooled together.

4.1.7 Clfdr

For the conditional local fdr (CLfdr) [20] procedure, hypotheses were stratified as for the GBH procedure. Within each bin the local fdrs (i. e., alternative densities and π_0) were estimated using the R package `fdrtool` [5]. Note that the authors of [20] suggested using a different estimator of the local fdr, namely the one in [51]. However, since the latter makes a Normal assumption and operates on z-scores rather than p-values, we opted for `fdrtool`.

4.1.8 FDRreg

For FDRreg [23] we used the implementation available on the first author's github site (`http://github.com/jgscott/FDRreg`). The FDRreg method has multiple tuning parameters. We used settings similar to the ones employed in these authors' study of neural synchrony detection [23]. In particular, the design matrix consisted of expanding the covariate in a cubic B-spline basis with 3 degrees of freedom. The regularization parameter λ for the ridge penalty was set to $\lambda = 0.01$. In addition, in contrast to the other methods benchmarked here, FDRreg operates on z-scores rather than p-values. To make comparison feasible, we converted each

p-value P into a z-score with the formula $Z = \Phi^{-1}(P)$, where Φ is the cumulative distribution of a standard Normal random variable. Also FDRreg estimates the conditional fdr in both tails of the z-score distribution, while the original p-values do not contain this two-tailed information. Therefore, in order to avoid spurious rejections caused by estimation on the right tail of the distribution, where p-values ≥ 0.5 get mapped, we set the local fdr of p-values ≥ 0.5 to 1. This makes the comparison more fair, since none of the other methods rejected p-values ≥ 0.5 in our simulation settings.

4.2 Simulations

For our experimental results (numerical simulations), we used three simulation scenarios:

4.2.1 All nulls simulation

For the all nulls simulations, we drew independent uniform random variables (for $i \in \{1, \ldots, m\}$):

$$\begin{aligned} P_i &\sim U[0,1] \\ X_i &\sim U[0,1] \\ H_i &= 0 \end{aligned}$$

We used $m = 20000$ and 4000 Monte Carlo replications. The methods were evaluated for nominal FDR control values $\alpha \in [0.01, 0.1]$ (equidistant grid with 10 values).

4.2.2 Effect size simulation

The two-sample t-test was applied to Normal samples ($n = 2 \times 5$, σ=1) with either the same mean (nulls) or means differing by the effect size ξ_i (alternatives). The fraction of true alternatives was 0.05. The pooled variance was used as the covariate.
We used $m = 20000$ and 1000 Monte Carlo replications. The simulations were repeated for 20 values of the simulation parameter $\xi \in [1, 2.5]$ (equidistant grid). The nominal α was set to 0.1.

4.2.3 Size investing simulation

For the size investing simulation simulations, we drew independent uniform random variables (for $i \in \{1, \ldots, m\}$):

$$\begin{aligned} H_i &\sim \text{Bernoulli}(\pi_1) \\ X_i &\sim U[1, \xi_{\max}] \\ Z_i &\sim \mathcal{N}(H_i X_i, 1) \\ P_i &= 1 - \Phi(Z_i) \end{aligned}$$

Φ denotes the standard Normal distribution function. P_i and X_i were used as the p-values and covariates respectively. We used $\pi_1 = 0.1$, $m = 20000$ and 500 Monte Carlo replicates. The simulations were repeated for 10 values of the simulation parameter $\xi_{\max} \in [3, 6]$ (equidistant grid). The nominal α was set to 0.1.

4.3 Evaluation of simulations

We used the following metric to evaluate the methods for power and FDR control. For a fixed multiple testing method, let $\Delta_i \in \{0, 1\}$ indicate whether the method rejected hypothesis i ($\Delta_i = 1$) or whether it did not ($\Delta_i = 0$). Then we defined power as

$$\text{Pow} = \mathbb{E}\left[\frac{\sum_{i=1}^{m} \Delta_i H_i}{1 \vee \sum_{i=1}^{m} H_i}\right]$$

and the FDR as

$$\text{FDR} = \mathbb{E}\left[\frac{\sum_{i=1}^{m} \Delta_i(1 - H_i)}{1 \vee \sum_{i=1}^{m} \Delta_i}\right]$$

Both of these quantities were estimated from their empirical counterparts based on the Monte Carlo replications.

Supplementary Note 5: tdr and size-investing

Here we illustrate how size-investing strategies can be derived in terms of tdr (equivalently in terms of fdr). For this, we will simplify things by reducing the multiple testing problem to that of testing two hypotheses H and $\widetilde{H}$ (the case of $m > 2$ hypotheses follows by just throwing in more notation).
Let P and $\widetilde{P}$ be the p-values for the two hypotheses and $F(t) = \pi_0 t + (1-\pi_0)F_1(t)$, $\widetilde{F}(t) = \widetilde{\pi_0}t + (1-\widetilde{\pi_0})\widetilde{F_1}(t)$ the corresponding distribution functions. We also assume that we have more power to detect H than $\widetilde{H}$, which we define as $F(t) > \widetilde{F}(t)\ \forall\ t$.
For an optimal procedure it would now hold that $fdr(t) = \widetilde{fdr}(\widetilde{t}) = \alpha$ [21] for whichever value of α is specified. Size-investing can now be explained as follows:

- For α small enough, because $F > \widetilde{F}$ it follows from $fdr(t) = \widetilde{fdr}(\widetilde{t}) = \alpha$ that $t \geq \widetilde{t}$, i.e. we assign more weight to H than to $\widetilde{H}$. In terms of size-investing, this corresponds to the case where our type-I error budget it small enough, that we to try to assign most of it to the more powerful hypothesis.

- On the other hand, as we increase our budget α, it is possible that eventually from $fdr(t) = \widetilde{fdr}(\widetilde{t}) = \alpha$ it will follow that $t \leq \widetilde{t}$, i.e. we assign higher weight to $\widetilde{H}$. This corresponds to the case where we have enough type-I error budget and enough power in H, that we should prioritize the test with the lower power, namely $\widetilde{H}$.

In other words, for size-investing to occur, the fdr and $\widetilde{fdr}$ curves have to cross. But in a particular case, this cannot happen, namely when $F_1 = \widetilde{F_1}$ and $F > \widetilde{F}$ is only due to $\pi_0 < \widetilde{\pi_0}$. In this case, $fdr(t) = \widetilde{fdr}(\widetilde{t}) = \alpha \Rightarrow t \geq \widetilde{t}$ for all values of α. As a consequence, any method which assumes that the alternative distribution is the same for all hypotheses and that only π_0 varies will not be able to apply a size-investing strategy.
A graphical explanation of these ideas is shown in Supplementary Figure 3.

Supplementary Note 6: IHW-Bonferroni for FWER control

In this section we show how the IHW ideas presented in the context of FDR control can be adapted in a straightforward way to Bonferroni's multiple testing procedure. This yields a new, powerful FWER controlling procedure, which we also implemented in the IHW package. We prove that this Bonferroni method with data-driven weights has finite sample FWER control.

6.1 Motivation and description

To extend the previous ideas to FWER control, we first quickly consider the classic approaches, which do not take covariate information into account. An analogon to the equivalence theorem (10) can in many cases be applied. In particular, many FWER controlling procedures can be interpreted as follows:
Reject all p-values $\leq t^*$, where t^* solves the optimization problem:

$$\text{maximize } R(t), \text{ s.t. } \widehat{\text{Fwer}}(t) \leq \alpha, \ t \in [0,1], \tag{34}$$

and $\widehat{\text{Fwer}}(t)$ is an appropriate conservative estimator of $\text{FWER}(t)$. In addition, things often are very simple, because $\widehat{\text{Fwer}}(t)$ is deterministic for many FWER controlling procedures. (In contrast, for FDR control, $\widehat{\text{Fdr}}(t)$ will be random, c. f. (9).) For example, assuming that the p-values are independent leads to an estimator $\widehat{\text{Fwer}}(t)$ that corresponds to the Šidák correction.
For the Bonferroni procedure the FWER is upper bounded by Markov's inequality,

$$\text{FWER}(t) = \Pr(V(t) \geq 1) \leq \mathbb{E}[V(t)] \leq mt. \tag{35}$$

In other words, it uses $\widehat{\text{Fwer}}(t) = mt$ in Equation (34). The optimization problem has the analytic solution $t^* = \frac{\alpha}{m}$.
From this it is clear how optimal weights/thresholds can be chosen in the stratified case, in analogy to (21) for FDR control:

$$\text{maximize } R(t_1, \ldots, t_G), \text{ s.t. } \sum_{g=1}^{G} m_g t_g \leq \alpha, \ t_g \in [0,1] \tag{36}$$

Modification (E1) is now readily applicable and as in (29), we get the convex optimization problem:

$$\begin{aligned} \text{minimize } \ H(t) &= -\sum_{g=1}^{G} m_g \widetilde{F_g}(t_g) \quad \text{s.t.} \\ H_1(t) &= \sum_{g=1}^{G} m_g t_g \leq \alpha \\ t_g &\in [0,1] \end{aligned} \tag{37}$$

After solving this optimization problem, we recover thresholds $t_1, \ldots, t_G$ which can be converted to weights via equation (24). Then we can apply the weighted Bonferroni procedure [6] with these weights. Modifications (E2) and (E3) are also immediately applicable. We call the resulting procedure **IHW-Bonferroni**.

6.2 Proof of finite sample FWER control

Theorem 2. *Let $H_i \in \{0,1\}$ be deterministic, $P_i \in [0,1]$ and $X_i \in \mathcal{X}$ random and assume that $(P_i, X_i), \quad i \in \{1,\dots,m\}$ are mutually independent. In addition, assume that $P_i \perp X_i \mid H_i = 0$ and that $P_i \sim U[0,1] \mid H_i = 0$. Then, the IHW-Bonferroni procedure controls the FWER at level α:*

$$\mathrm{FWER}_{\mathrm{IHW-Bonferroni}} \leq \alpha$$

Remark 2*:* Here, for simplicity of the proof we assume that the hypotheses are deterministic. The random case can be handled similarly after assuming that $(P_i, X_i, H_i), \quad i \in \{1,\dots,m\}$ are mutually independent.

Proof. First we will show that $P_i \perp W_i \mid H_i = 0$. For this let i such that $H_i = 0$ and assume that hypothesis i gets assigned to fold l during step (E2) of IHW-Bonferroni. Also denote by $\mathbf{P}_{-l}$, resp. $\mathbf{X}_{-l}$ the vector of p-values, resp. covariates that were not assigned to fold l. Also let L be the cardinality of these vectors. Notice that step (E2) provides a measurable function $h_l : \mathcal{X} \times \mathcal{X}^L \times [0,1]^L$, such that:

$$W_i = h_l(X_i, \mathbf{X}_{-l}, \mathbf{P}_{-l})$$

But by the assumptions of the theorem and because $H_i = 0$, it follows that P_i is independent of $(X_i, \mathbf{X}_{-l}, \mathbf{P}_{-l})$. Therefore it also follows that $P_i \perp W_i$ and by choice of i we get our desired statement. We can now conclude our proof as follows:

$$\begin{aligned}
\mathrm{FWER} &= \Pr\left(V \geq 1\right) \\
&= \Pr\left(\bigcup_{i \in \mathscr{H}_0} \left\{Q_i \leq \frac{\alpha}{m}\right\}\right) \\
&= \Pr\left(\bigcup_{i \in \mathscr{H}_0} \left\{P_i \leq \frac{\alpha W_i}{m}\right\}\right) \\
&\leq \sum_{i \in \mathscr{H}_0} \Pr\left(P_i \leq \frac{\alpha W_i}{m}\right) \\
&= \sum_{i \in \mathscr{H}_0} \mathbb{E}\left[\Pr\left(P_i \leq \frac{\alpha W_i}{m} \middle| W_i\right)\right] \\
&= \sum_{i \in \mathscr{H}_0} \mathbb{E}\left[\frac{\alpha W_i}{m}\right] \quad (\text{since } P_i \sim U[0,1], P_i \perp W_i \text{ for } i \in \mathscr{H}_0) \\
&\leq \frac{\alpha}{m} \sum_{i=1}^{m} \mathbb{E}\left[W_i\right] \\
&= \frac{\alpha}{m} \mathbb{E}\left[\sum_{i=1}^{m} W_i\right] \\
&\leq \alpha
\end{aligned}$$

□

Remark 3*:* Note that the proof can be generalized. Rather than requiring joint independence of $(P_i, X_i, H_i), \quad i \in \{1,\dots,m\}$ it is already sufficient if $(\mathbf{P}_l, \mathbf{X}_l), \quad l \in \{1,\dots,n_{\mathrm{folds}}\}$ are mutually independent, where the vectors $\mathbf{P}_l, \mathbf{X}_l$ correspond to hypotheses assigned to the l-th fold. This implies that IHW and IHW-Bonferroni could be easily extended to account for this dependence structure (when it is known), rather than randomly assigning folds as in (E2).

6.3 Empirical performance of IHW-Bonferroni

We further demonstrate the theoretical results in Subsection 6.2 using the same simulation scenarios as described for IHW in Supplementary Note 4. In particular, we benchmark IHW-Bonferroni (with settings as described in Subsubsection 4.1.2, where the nominal FDR control level α is instead the nominal FWER control level) against Bonferroni (`p.adjust` function with method `bonferroni` in R's `stats` package).
The simulation results are shown in Supplementary Figure 5. Under all three simulations scenarios, both Bonferroni and IHW-Bonferroni control the FWER, as theoretically expected. In addition, IHW-Bonferroni increases power compared to Bonferroni; the improvement can be dramatic.

Supplementary Note 7: Proof of asymptotic consistency of IHW

In this section, we want to prove that IHW is asymptotically consistent, i. e., that IHW controls the FDR at the nominal level α as the number of hypotheses becomes large. The main result is stated in Theorem 4. To make the proof more readable, we use the simplified setup already used for the presentation of the algorithm in Section 2. In particular, we assume that we have a discrete covariate that takes on a finite number of levels $g = 1, \ldots, G$, and to be more precise we consider the grouped setting as in [10, 46]. This is also how IHW is operationalized for continuous covariates by means of stratification.

Thus, assume that we have G strata and in the g^{th} stratum we have m_g p-values $P_1^g, \ldots, P_{m_g}^g$, $g = 1, \ldots, G$. Of these, $m_{0,g}$ correspond to null hypotheses and $m_g - m_{0,g}$ to alternatives. The setup we consider is conditional on the true status H_i^g of each hypothesis. Given a threshold $t_g \in [0, 1]$ we write $V_g(t_g) = \sum_{i=1}^{m_g} \mathbf{1}_{\{P_i^g \leq t_g, H_i^g = 0\}}$ for the number of falsely rejected hypotheses and $R_g(t_g) = \sum_{i=1}^{m_g} \mathbf{1}_{\{P_i^g \leq t_g\}}$ for the number of all rejected hypotheses $P_i^g \leq t_g$ in stratum g.

We also assume that the two-groups model holds within each stratum as follows:

$$P_i^g \mid H_i^g = 0 \sim U[0, 1]$$
$$P_i^g \mid H_i^g = 1 \sim F_{1,g}$$

Note that the uniform distribution under the null is implied by the conditional independence of covariates and p-values under the null, while the alternative distribution $F_{1,g}$, $g = 1, \ldots, G$ will in general depend on g, because the covariate would have been selected to be associated with power.

For our proofs, we will also require the following assumptions:

Assumption 1. *The distribution under the alternative $F_{1,g}$ is continuous $\forall g \in \{1, \ldots, G\}$ and $F_{1,g}(t) > 0 \; \forall \, t \in (0, 1]$.*

Assumption 2.

$$\forall \, g \in \{1, \ldots, G\} : \quad \frac{m_g}{m} \to \tilde{\pi}_g \text{ as } m \to \infty, \quad \tilde{\pi}_g \in (0, 1)$$

Assumption 3.

$$\forall \, g \in \{1, \ldots, G\} : \quad \frac{m_{0,g}}{m_g} \to \pi_{0,g} \text{ as } m \to \infty, \quad \pi_{0,g} \in (0, 1)$$

Assumption 4. *The hypotheses within each group satisfy the weak dependence criterion, in other words for all $g \in \{1, \ldots, G\}$, the following holds:*

1. $\dfrac{V_g(t)}{m_{0,g}} \to t, \quad m \to \infty$
2. $\dfrac{R_g(t) - V_g(t)}{m_g - m_{0,g}} \to F_{1,g}(t), \quad m \to \infty$

almost surely for each $t \in (0, 1]$.

Remark 4: The above assumptions are standard for proving asymptotics for FDR control. They are very similar to the assumptions in [3], extended to the grouped setting. They are also similar to the assumptions made in [10]. Our proof closely follows the techniques in [3], the results of which we generalize to the grouped setting.

Remark 5*:* In Assumption 2 we introduced $\tilde{\pi}_g$. Intuitively, $\tilde{\pi}_g$ is the prior probability of a hypothesis belonging to stratum g.

We quickly repeat the most important notation, most of which has already been introduced above and in Section 2:

- $V(t_1,\dots,t_G) = \sum_{g=1}^{G} V_g(t_g) = \sum_{g=1}^{G}\sum_{i=1}^{m_g} \mathbf{1}_{\{P_i^g \le t_g, H_i^g = 0\}}$
- $R(t_1,\dots,t_G) = \sum_{g=1}^{G} R_g(t_g) = \sum_{g=1}^{G}\sum_{i=1}^{m_g} \mathbf{1}_{\{P_i^g \le t_g\}}$
- $\widehat{\mathrm{Fdr}}(t_1,\dots,t_G) = \dfrac{\sum_{g=1}^{G} m_g t_g}{R(t_1,\dots,t_g)\vee 1}$
- $\mathrm{FDR}(t_1,\dots,t_G) = \mathbb{E}\left[\dfrac{V(t_1,\dots,t_g)}{R(t_1,\dots,t_g)\vee 1}\right]$

We also introduce the following notation:

- $F(t_1,\dots,t_G) = \sum_{g=1}^{G} \tilde{\pi}_g \left[\pi_{0,g} t_g + (1-\pi_{0,g}) F_{1,g}(t_g)\right]$
- $\mathrm{Fdr}^0(t_1,\dots,t_G) = \dfrac{\sum_{g=1}^{G} \tilde{\pi}_g t_g}{F(t_1,\dots,t_G)}$

Also, instead of the parametrization $(t_1,\dots,t_G)$ we will sometimes parametrize the above functions by $(t,\mathbf{w})$, where $\mathbf{w}\in\mathbb{R}^G_{\ge 0}$. The parametrization will be given by $(t,\mathbf{w}) \mapsto (w_1 t \wedge 1,\dots,w_G t\wedge 1)$, and for example we will write $F(t,\mathbf{w})$ or $R(t,\mathbf{w})$ for the induced functions. Note that for notational convenience we will write e. g., $w_g t$ instead of $w_g t\wedge 1$ whenever the truncation is obvious by the context.

We also introduce notation for the Euclidean subsets in which the weight vectors lie:

Definition 1. *(1)* $\Delta^G := \left\{\mathbf{w}\in\mathbb{R}^G \ : \ w_g \ge 0, \sum_{g=1}^{G} \tilde{\pi}_g w_g = 1\right\}$

(2) $\Delta^G_\varepsilon := \left\{\mathbf{w}\in\mathbb{R}^G \ : \ w_g \ge 0, \left|\sum_{g=1}^{G} \tilde{\pi}_g w_g - 1\right| \le \varepsilon\right\}$ *for* $\varepsilon > 0$

Lemma 1. *Let Assumptions 1-4 hold and let* $\frac{1}{2} > \varepsilon > 0$*. Then it follows that:*

(1) $\sup_{\substack{\mathbf{w}\in\Delta^G_\varepsilon \\ t\in[0,1]}} \left|\dfrac{V(t,\mathbf{w})}{m} - \sum_{g=1}^{G} \tilde{\pi}_g \pi_{0,g} w_g t\right| \to 0$ *almost surely as* $m\to\infty$

(2) $\sup_{\substack{\mathbf{w}\in\Delta^G_\varepsilon \\ t\in[0,1]}} \left|\dfrac{R(t,\mathbf{w})}{m} - F(t,\mathbf{w})\right| \to 0$ *almost surely as* $m\to\infty$

Proof. We start with the proof of (1). Note that as in [3], the Assumptions 3 and 4 yield, by a simple modification of the proof of Glivenko-Cantelli [52], that:

$$\sup_{t_g \in [0,1]} \left| \frac{V_g(t_g)}{m_g} - \pi_{0,g} t_g \right| \to 0 \quad \text{almost surely as } m \to \infty \quad \forall\, g \in \{1, \dots, G\}$$

Based on the above, we get:

$$\begin{aligned}
&\sup_{\substack{\mathbf{w} \in \Delta_\varepsilon^G \\ t \in [0,1]}} \left| \frac{V(t, \mathbf{w})}{m} - \sum_{g=1}^{G} \tilde{\pi}_g \pi_{0,g} w_g t \right| \\
\leq &\sup_{t_1, \dots, t_G \in [0,1]} \left| \frac{V(t_1, \dots, t_G)}{m} - \sum_{g=1}^{G} \tilde{\pi}_g \pi_{0,g} t_g \right| \\
= &\sup_{t_1, \dots, t_G \in [0,1]} \left| \frac{\sum_{g=1}^{G} V_g(t_g)}{m} - \sum_{g=1}^{G} \tilde{\pi}_g \pi_{0,g} t_g \right| \\
\leq &\sum_{g=1}^{G} \sup_{t_g \in [0,1]} \left| \frac{m_g}{m} \frac{V_g(t_g)}{m_g} - \tilde{\pi}_g \pi_{0,g} t_g \right| \\
\leq &\sum_{g=1}^{G} \left[\tilde{\pi}_g \sup_{t_g \in [0,1]} \left| \frac{V_g(t_g)}{m_g} - \pi_{0,g} t_g \right| + \left| \tilde{\pi}_g - \frac{m_g}{m} \right| \right] \\
\to &\, 0 \quad \text{almost surely as } m \to \infty
\end{aligned}$$

The convergence follows by Assumption 2 and the Glivenko-Cantelli type results above.

In the same way we can show the analogous result for the hypotheses under the alternative and then another application of the triangle inequality yields (2). □

Lemma 2. *Let* $1 \geq \delta > 0$, $\frac{1}{2} > \varepsilon > 0$ *and let Assumptions 1-4 hold. Then it follows that:*

$$\sup_{\substack{\mathbf{w} \in \Delta_\varepsilon^G \\ t \in [\delta, 1]}} \left| \widehat{\mathrm{Fdr}}(t, \mathbf{w}) - \mathrm{Fdr}^0(t, \mathbf{w}) \right| \to 0 \quad \textit{almost surely as } m \to \infty$$

Proof.

$$
\begin{aligned}
&\sup_{\substack{\mathbf{w}\in\Delta_\varepsilon^G\\ t\in[\delta,1]}}\left|\widehat{\mathrm{Fdr}}(t,\mathbf{w})-\mathrm{Fdr}^0(t,\mathbf{w})\right|\\
=&\sup_{\substack{\mathbf{w}\in\Delta_\varepsilon^G\\ t\in[\delta,1]}}\left|\frac{\sum_{g=1}^G m_g w_g t}{R(t,\mathbf{w})\vee 1}-\frac{\sum_{g=1}^G \tilde{\pi}_g w_g t}{F(t,\mathbf{w})}\right|\\
\leq&\sup_{\substack{\mathbf{w}\in\Delta_\varepsilon^G\\ t\in[\delta,1]}}\left|\frac{\sum_{g=1}^G(\frac{m_g}{m}-\tilde{\pi}_g)w_g t}{\frac{R(t,\mathbf{w})\vee 1}{m}}+\frac{\sum_{g=1}^G \tilde{\pi}_g w_g t}{\frac{R(t,\mathbf{w})\vee 1}{m}}-\frac{\sum_{g=1}^G \tilde{\pi}_g w_g t}{F(t,\mathbf{w})}\right|\\
\leq&\frac{\sum_{g=1}^G\left|\frac{m_g}{m}-\tilde{\pi}_g\right|}{\inf\limits_{\substack{\mathbf{w}\in\Delta_\varepsilon^G\\ t\in[\delta,1]}}\frac{R(t,\mathbf{w})\vee 1}{m}}+\frac{\sup\limits_{\substack{\mathbf{w}\in\Delta_\varepsilon^G\\ t\in[0,1]}}\sum\limits_{g=1}^G\tilde{\pi}_g w_g t\sup\limits_{\substack{\mathbf{w}\in\Delta_\varepsilon^G\\ t\in[0,1]}}\left|\frac{R(t,\mathbf{w})\vee 1}{m}-F(t,\mathbf{w})\right|}{\inf\limits_{\substack{\mathbf{w}\in\Delta_\varepsilon^G\\ t\in[\delta,1]}}F(t,\mathbf{w})\inf\limits_{\substack{\mathbf{w}\in\Delta_\varepsilon^G\\ t\in[\delta,1]}}\frac{R(t,\mathbf{w})\vee 1}{m}}\\
\leq&\frac{\sum_{g=1}^G\left|\frac{m_g}{m}-\tilde{\pi}_g\right|}{\inf\limits_{\mathbf{w}\in\Delta_\varepsilon^G}\frac{R(\delta,\mathbf{w})\vee 1}{m}}+\frac{(1+\varepsilon)\sup\limits_{\substack{\mathbf{w}\in\Delta_\varepsilon^G\\ t\in[0,1]}}\left|\frac{R(t,\mathbf{w})\vee 1}{m}-F(t,\mathbf{w})\right|}{\inf\limits_{\mathbf{w}\in\Delta_\varepsilon^G}F(\delta,\mathbf{w})\inf\limits_{\mathbf{w}\in\Delta_\varepsilon^G}\frac{R(\delta,\mathbf{w})\vee 1}{m}}\\
\to&\,0\quad\text{almost surely as } m\to\infty
\end{aligned}
$$

In this case, the convergence follows by Assumption 2, Lemma 1 and because $\inf\limits_{\mathbf{w}\in\Delta_\varepsilon^G}F(\delta,\mathbf{w})>0$. The latter follows because of the compactness of Δ_ε^G, the continuity of F and because:

$$F(\delta,\mathbf{w})>0\;\forall\;\mathbf{w}\in\Delta_\varepsilon^G$$

□

Theorem 3. *Let $1\geq\delta>0$, $\frac{1}{2}>\varepsilon>0$ and let Assumptions 1-4 hold. Then:*

$$\liminf_{m\to\infty}\inf_{\substack{\mathbf{w}\in\Delta_\varepsilon^G\\ t\in[\delta,1]}}\left\{\widehat{\mathrm{Fdr}}(t,\mathbf{w})-\frac{V(t,\mathbf{w})}{R(t,\mathbf{w})\vee 1}\right\}\geq 0\quad\textit{almost surely}\tag{38}$$

Proof. First note that:

$$
\begin{aligned}
&\sup_{\substack{\mathbf{w}\in\Delta_\varepsilon^G\\ t\in[\delta,1]}}\left|\frac{\sum_{g=1}^G\tilde{\pi}_g\pi_{0,g}w_g t}{\frac{R(t,\mathbf{w})\vee 1}{m}}-\frac{V(t,\mathbf{w})}{R(t,\mathbf{w})\vee 1}\right|\\
\leq&\frac{1}{\inf\limits_{\substack{\mathbf{w}\in\Delta_\varepsilon^G\\ t\in[\delta,1]}}\frac{R(t,\mathbf{w})\vee 1}{m}}\sup_{\substack{\mathbf{w}\in\Delta_\varepsilon^G\\ t\in[0,1]}}\left|\sum_{g=1}^G\tilde{\pi}_g\pi_{0,g}w_g t-\frac{V(t,\mathbf{w})}{m}\right|\\
\to&\,0\quad\text{almost surely as } m\to\infty
\end{aligned}
$$

This follows by Lemma 1 and a similar argument as in the proof of Lemma 2. In addition, by construction it holds that for all $\mathbf{w}\in\mathbb{R}^G_{\geq 0}$:

$$
\begin{aligned}
&\liminf_{m\to\infty}\left\{\widehat{\mathrm{Fdr}}(t,\mathbf{w})-\frac{\sum_{g=1}^{G}\tilde{\pi}_g\pi_{0,g}w_g t}{\frac{R(t,\mathbf{w})\vee 1}{m}}\right\}\\
=&\liminf_{m\to\infty}\left\{\frac{\sum_{g=1}^{G}\frac{m_g}{m}w_g t}{\frac{R(t,\mathbf{w})\vee 1}{m}}-\frac{\sum_{g=1}^{G}\tilde{\pi}_g\pi_{0,g}w_g t}{\frac{R(t,\mathbf{w})\vee 1}{m}}\right\}\\
=&\liminf_{m\to\infty}\left\{\frac{\sum_{g=1}^{G}\left(\frac{m_g}{m}-\tilde{\pi}_g\pi_{0,g}\right)w_g t}{\frac{R(t,\mathbf{w})\vee 1}{m}}\right\}\geq 0 \text{ almost surely}
\end{aligned}
$$

The last inequality holds, because by Assumption 2: $\frac{m_g}{m}\to\tilde{\pi}_g\geq\tilde{\pi}_g\pi_{0,g}$. Combining the two results above, yields the wanted result:

$$
\liminf_{m\to\infty}\inf_{\substack{\mathbf{w}\in\Delta_\varepsilon^G\\ t\in[\delta,1]}}\left\{\widehat{\mathrm{Fdr}}(t,\mathbf{w})-\frac{V(t,\mathbf{w})}{R(t,\mathbf{w})\vee 1}\right\}\geq 0 \quad \text{almost surely}
$$

□

For the final proof of asymptotic consistency of IHW, we will need one additional assumption:

Assumption 5. *There exists* $t'\in(0,1]$ *such that:* $\sup_{\mathbf{w}\in\Delta^G}\mathrm{Fdr}^0(t',\mathbf{w})<\alpha$.

Theorem 4. *Let Assumptions 1-5 hold. Also let* $\mathbf{w}^*$ *be a (possibly data-driven) weight vector, i.e. satisfy* $w_g^*\geq 0$ *and* $\sum_{g=1}^G m_g w_g^*=m$. *Define:*

$$
t^*(\mathbf{w}^*)=\sup_{t\in[0,1]}\left\{t\,|\,\widehat{\mathrm{Fdr}}(t,\mathbf{w}^*)\leq\alpha\right\}
$$

Then the weighted Benjamini-Hochberg procedure with (data-driven) weights $\mathbf{w}^*$ *(assigned by a measurable rule) asymptotically controls the FDR, in other words:*

$$
\limsup_{m\to\infty}\mathrm{FDR}(t^*(\mathbf{w}^*),\mathbf{w}^*)\leq\alpha
$$

Proof. Let t' be as in Assumption 5, i.e. assume that $\alpha-\sup_{\mathbf{w}\in\Delta^G}\mathrm{Fdr}^0(t',\mathbf{w})>0$. By continuity of Fdr^0, there exists $\varepsilon>0$ such that $\alpha-\sup_{\mathbf{w}\in\Delta_\varepsilon^G}\mathrm{Fdr}^0(t',\mathbf{w})=:\tilde{\varepsilon}>0$. For this ε, pick M large enough, such that $\mathbf{w}^*\in\Delta_\varepsilon^G\ \forall m\geq M$. Such M exists by Assumption 2.

By Lemma 2, we get that for almost all ω in our sample space, there exists $\widetilde{M}(\omega)$ large enough, such that for all $m\geq\widetilde{M}(\omega)$:

$$
\left|\widehat{\mathrm{Fdr}}(t',\mathbf{w}^*)-\mathrm{Fdr}^0(t',\mathbf{w}^*)\right|\leq\frac{\tilde{\varepsilon}}{2}
$$

Therefore it follows that: $\widehat{\mathrm{Fdr}}(t',\mathbf{w}^*)<\alpha$ and thus: $t^*(\mathbf{w}^*)\geq t'$.

In particular we get:

$$
\liminf_{m\to\infty}t^*(\mathbf{w}^*)\geq t' \text{ almost surely}
$$

Then let $\delta = \frac{t'}{2} > 0$. By Theorem 3 it holds with probability 1 that:

$$\liminf_{m\to\infty}\left[\widehat{\mathrm{Fdr}}(t^*(\mathbf{w}^*),\mathbf{w}^*) - \frac{V(t^*(\mathbf{w}^*),\mathbf{w}^*)}{R(t^*(\mathbf{w}^*),\mathbf{w}^*)\vee 1}\right] \geq \liminf_{m\to\infty}\inf_{\substack{\mathbf{w}\in\Delta_\varepsilon^G\\ t\in[\delta,1]}}\left\{\widehat{\mathrm{Fdr}}(t,\mathbf{w}) - \frac{V(t,\mathbf{w})}{R(t,\mathbf{w})\vee 1}\right\} \geq 0$$

In particular, since by definition $\widehat{\mathrm{Fdr}}(t^*(\mathbf{w}^*),\mathbf{w}^*) \leq \alpha$, it also follows with probability 1 that:

$$\limsup_{m\to\infty}\frac{V(t^*(\mathbf{w}^*),\mathbf{w}^*)}{R(t^*(\mathbf{w}^*),\mathbf{w}^*)\vee 1} \leq \alpha$$

Applying the reverse Fatou Lemma yields the result, i.e.:

$$\limsup_{m\to\infty}\mathrm{FDR}(t^*(\mathbf{w}^*),\mathbf{w}^*) = \limsup_{m\to\infty}\mathbb{E}\left[\frac{V(t^*(\mathbf{w}^*),\mathbf{w}^*)}{R(t^*(\mathbf{w}^*),\mathbf{w}^*)\vee 1}\right] \leq \mathbb{E}\left[\limsup_{m\to\infty}\frac{V(t^*(\mathbf{w}^*),\mathbf{w}^*)}{R(t^*(\mathbf{w}^*),\mathbf{w}^*)\vee 1}\right] \leq \alpha$$

□

Remark 6*:* Thereom 4 implies that IHW with modification (E1) asymptotically controls the FDR. Similarly, IHW with modifications (E1) and (E3) also asymptotically controls the FDR. Note that at least intuitively (and shown in our simulations), modifications (E2) and (E3) make our procedure even more conservative. The full procedure is also asymptotically consistent; but to include modification (E2), the proof needs a straightforward modification: Instead of considering G groups based on the covariate, we need to consider the $G \times n_{\text{folds}}$ groups as generated by all combinations of covariate levels and folds. Then the proof proceeds in exactly the same fashion.

References

[1] Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* 289–300 (1995).

[2] Benjamini, Y., Krieger, A. M. & Yekutieli, D. Adaptive linear step-up procedures that control the false discovery rate. *Biometrika* **93**, 491–507 (2006).

[3] Storey, J. D., Taylor, J. E. & Siegmund, D. Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: a unified approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **66**, 187–205 (2004).

[4] Efron, B. *Large-scale inference: Empirical Bayes methods for estimation, testing, and prediction* (Cambridge University Press, 2010).

[5] Strimmer, K. A unified approach to false discovery rate estimation. *BMC Bioinformatics* **9**, 303 (2008).

[6] Genovese, C. R., Roeder, K. & Wasserman, L. False discovery control with p-value weighting. *Biometrika* **93**, 509–524 (2006).

[7] Roeder, K., Devlin, B. & Wasserman, L. Improving power in genome-wide association studies: weights tip the scale. *Genetic Epidemiology* **31**, 741–747 (2007).

[8] Roquain, E. & Van De Wiel, M. Optimal weighting for false discovery rate control. *Electronic Journal of Statistics* **3**, 678–711 (2009).

[9] Bourgon, R., Gentleman, R. & Huber, W. Independent filtering increases detection power for high-throughput experiments. *Proceedings of the National Academy of Sciences* **107**, 9546–9551 (2010).

[10] Hu, J. X., Zhao, H. & Zhou, H. H. False discovery rate control with groups. *Journal of the American Statistical Association* **105** (2010).

[11] Dobriban, E., Fortney, K., Kim, S. K. & Owen, A. B. Optimal multiple testing under a Gaussian prior on the effect sizes. *Biometrika* **102**, 753–766 (2015).

[12] Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-Seq data with DESeq2. *Genome Biology* **15**, 550 (2014).

[13] Bottomly, D. *et al.* Evaluating gene expression in C57BL/6J and DBA/2J mouse striatum using RNA-Seq and microarrays. *PLoS ONE* **6**, e17820 (2011).

[14] Frazee, A. C., Langmead, B. & Leek, J. T. Recount: a multi-experiment resource of analysis-ready rna-seq gene count datasets. *BMC Bioinformatics* **12**, 449 (2011).

[15] Dephoure, N. & Gygi, S. P. Hyperplexing: a method for higher-order multiplexed quantitative proteomics provides a map of the dynamic response to rapamycin in yeast. *Science Signaling* **5**, rs2–rs2 (2012).

[16] Grubert, F. *et al.* Genetic control of chromatin states in humans involves local and distal chromosomal interactions. *Cell* **162**, 1051–1065 (2015).

[17] Peña, E. A., Habiger, J. D. & Wu, W. Power-enhanced multiple decision functions controlling family-wise error and false discovery rates. *The Annals of Statistics* **39**, 556–583 (2011).

[18] Sun, W. & Cai, T. T. Oracle and adaptive compound decision rules for false discovery rate control. *Journal of the American Statistical Association* **102**, 901–912 (2007).

[19] Stephens, M. False discovery rates: A new deal. *bioRxiv* 038216 (2016).

[20] Cai, T. T. & Sun, W. Simultaneous testing of grouped hypotheses: Finding needles in multiple haystacks. *Journal of the American Statistical Association* **104** (2009).

[21] Ochoa, A., Storey, J. D., Llinás, M. & Singh, M. Beyond the E-value: Stratified statistics for protein domain prediction. *PLoS Computational Biology* **11**, e1004509 (2015).

[22] Ploner, A., Calza, S., Gusnanto, A. & Pawitan, Y. Multidimensional local false discovery rate for microarray studies. *Bioinformatics* **22**, 556–565 (2006).

[23] Scott, J. G., Kelly, R. C., Smith, M. A., Zhou, P. & Kass, R. E. False discovery rate regression: an application to neural synchrony detection in primary visual cortex. *Journal of the American Statistical Association* **110**, 459–471 (2015).

[24] Ferkingstad, E., Frigessi, A., Rue, H., Thorleifsson, G. & Kong, A. Unsupervised empirical Bayesian multiple testing with external covariates. *The Annals of Applied Statistics* 714–735 (2008).

[25] Efron, B. & Zhang, N. R. False discovery rates and copy number variation. *Biometrika* **98**, 251–271 (2011).

[26] Du, L. & Zhang, C. Single-index modulated multiple testing. *The Annals of Statistics* **42**, 30–79 (2014).

[27] Benjamini, Y. & Yekutieli, D. The control of the false discovery rate in multiple testing under dependency. *The Annals of Statistics* 1165–1188 (2001).

[28] Yoo, Y. J., Bull, S. B., Paterson, A. D., Waggott, D. & Sun, L. Were genome-wide linkage studies a waste of time? Exploiting candidate regions within genome-wide association studies. *Genetic Epidemiology* **34**, 107–118 (2010).

[29] Benjamini, Y. Discovering the false discovery rate. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **72**, 405–416 (2010).

[30] Storey, J. D. The positive false discovery rate: A Bayesian interpretation and the q-value. *The Annals of Statistics* 2013–2035 (2003).

[31] Efron, B. Microarrays, empirical Bayes and the two-groups model. *Statistical Science* **23**, 1–22 (2008).

[32] McClintick, J. N. & Edenberg, H. J. Effects of filtering by present call on analysis of microarray experiments. *BMC Bioinformatics* **7**, 49 (2006).

[33] Talloen, W. *et al.* I/NI-calls for the exclusion of non-informative genes: a highly effective filtering tool for microarray data. *Bioinformatics* **23**, 2897–2902 (2007).

[34] Hackstadt, A. J. & Hess, A. M. Filtering for increased power for microarray data analysis. *BMC Bioinformatics* **10**, 11 (2009).

[35] Benjamini, Y. Comment: Microarrays, empirical Bayes and the two-groups model. *Statistical Science* **23**, 23–28 (2008).

[36] Simmons, J. P., Nelson, L. D. & Simonsohn, U. False-positive psychology undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological science* 0956797611417632 (2011).

[37] Degner, J. F. *et al.* DNaseI sensitivity QTLs are a major determinant of human expression variation. *Nature* **482**, 390–394 (2012).

[38] Sun, L., Craiu, R. V., Paterson, A. D. & Bull, S. B. Stratified false discovery control for large-scale hypothesis testing with application to genome-wide association studies. *Genetic Epidemiology* **30**, 519–530 (2006).

[39] Efron, B. Simultaneous inference: When should hypothesis testing problems be combined? *The Annals of Applied Statistics* 197–223 (2008).

[40] Roeder, K. & Wasserman, L. Genome-wide significance levels and weighted hypothesis testing. *Statistical Science* **24**, 398 (2009).

[41] Roeder, K., Bacanu, S.-A., Wasserman, L. & Devlin, B. Using linkage genome scans to improve power of association in genome scans. *The American Journal of Human Genetics* **78**, 243–252 (2006).

[42] Li, L. *et al.* Using eQTL weights to improve power for genome-wide association studies: a genetic study of childhood asthma. *Frontiers in Genetics* **4** (2013).

[43] Poisson, L. M., Sreekumar, A., Chinnaiyan, A. M. & Ghosh, D. Pathway-directed weighted testing procedures for the integrative analysis of gene expression and metabolomic data. *Genomics* **99**, 265–274 (2012).

[44] Xing, C., Cohen, J. C. & Boerwinkle, E. A weighted false discovery rate control procedure reveals alleles at FOXA2 that influence fasting glucose levels. *The American Journal of Human Genetics* **86**, 440–446 (2010).

[45] Benjamini, Y. & Hochberg, Y. On the adaptive control of the false discovery rate in multiple testing with independent statistics. *Journal of Educational and Behavioral Statistics* **25**, 60–83 (2000).

[46] Zhao, H. & Zhang, J. Weighted p–value procedures for controlling fdr of grouped hypotheses. *Journal of Statistical Planning and Inference* (2014).

[47] Zablocki, R. W. *et al.* Covariate-modulated local false discovery rate for genome-wide association studies. *Bioinformatics* btu145 (2014).

[48] Shabalin, A. A. Matrix eQTL: ultra fast eQTL analysis via large matrix operations. *Bioinformatics* **28**, 1353–1358 (2012).

[49] Gurobi Optimization, I. Gurobi optimizer reference manual (2015).

[50] Lougee-Heimer, R. The common optimization interface for operations research: Promoting open-source software in the operations research community. *IBM Journal of Research and Development* **47**, 57–66 (2003).

[51] Jin, J. & Cai, T. T. Estimating the null and the proportion of nonnull effects in large-scale multiple comparisons. *Journal of the American Statistical Association* **102**, 495–506 (2007).

[52] van der Vaart, A. *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilistic Mathematics (Cambridge University Press, 2000).